# Who's A Good Decision Maker? Data-Driven Expert Worker Ranking under Unobservable Quality

*Completed Research Paper*

**Tomer Geva**
Coller School of Management
Tel Aviv University
Tel Aviv, Israel
tgeva@tau.ac.il

**Maytal Saar-Tsechansky**
McCombs School of Business
University of Texas at Austin
Austin, Texas
maytal@mail.utexas.edu

## Abstract

*Evaluation of expert workers by their decision quality has substantial practical value, yet using other expert workers for decision quality evaluation tasks is costly and often infeasible. In this work, we frame the Ranking of Expert workers according to their unobserved decision Quality (REQ) -- without resorting to evaluation by other experts -- as a new Data Science problem. This problem is challenging, as the correct decisions are commonly unobservable and substantial parts of the information available to the decision maker is not available for retrospective decision evaluation. We propose a new machine learning approach to address this problem. We evaluate our method on one dataset representing real expert decisions and two public datasets, and find that our approach is successful in generating highly accurate rankings. Moreover, we observe that our approach's superiority over the baseline is particularly prominent as evaluation settings become increasingly challenging.*

**Keywords:** Predictive Modeling, Decision Evaluation, Supervised Learning, Worker Ranking

## Introduction

In contemporary work environments, a significant portion of workers routinely make decisions of various levels of complexities. These include tasks with significant implications, such as medical diagnosis, financial fraud detection, and customer support diagnostics. Accurate ranking of expert workers by the quality of their decisions is key for successfully accomplishing organizational goals. In particular, such ranking is pivotal for informed incentive schemes, and for worker training and allocation decisions so as to enhance overall business performance. In some important settings, such as healthcare and security, the capacity to identify workers who are prone to decision errors also has significant societal implications.

Ranking professional workers by their decision quality is a challenging task because in many settings the correct decisions may remain unknown even after a decision is made. For example, physicians may determine a diagnosis and initiate treatment, while the true diagnosis may never be confirmed. Auditors of financial fraud or of healthcare fraud similarly make determinations of non-compliance or fraud, whereas failure to detect fraudulent claims may never be revealed. Consequently, it is not possible to directly compute the frequency of an expert's correct decisions. Furthermore, because experts often make non-trivial decisions, acquiring the correct decision entails having other highly paid experts assess the decision after the fact. For example, the opinions of expert physicians may be obtained at a significant cost to assess the quality of a given physician's diagnosis. Not only is such a process economically prohibitive in many settings, but ex-post peer assessments may often be socially or professionally unacceptable. Lastly, a complete record of the information that was available to the decision maker may not be available after the decision is made; thus, accurate and fair retrospective assessment by peer experts may be difficult or impossible.

In this paper, we first propose the new problem of Ranking Expert decision-makers' unobserved Quality (REQ): given multiple expert workers, where each worker routinely makes a large number decisions of unobserved quality and where workers encounter similar distribution of decision tasks – *how can we rank workers by the relative quality of their decisions without resorting to the acquisition of additional and potentially costly, peer-review by other experts?*

We then address this challenge by developing a data-driven approach to generate expert worker ranking. Our approach does not require *posteriori* knowledge of the correct decisions, nor access to complete knowledge of the information set on which workers based their decisions. Instead, our approach induces a predictive model to generate Pseudo Ground-Truth (PGT) decisions that correspond to the decisions inferred by a model. Workers are subsequently ranked based on the similarity of their decisions to the PGT decision generated by the model, while taking into account the confidence (or uncertainty) of the model's inference of the correct decision for each instance. As we will see, the confidence element of our approach is highly important. Especially, as correct decisions become more difficult to infer and when available data is more limited than the information available to the original decision makers.

We evaluated our approach using a real world dataset on sales tax auditor decisions, and repeated the analysis for robustness on two publically available datasets. Our evaluations employed simulations, controlling for workers' decision quality and distribution. We find that our REQ approach is highly effective for ranking workers by their decision quality, and that it yields highly accurate ranking even when the underlying mapping between the available predictors and the correct decisions are difficult for supervised models to induce, such as when the information used by the decision makers to arrive at their decisions is only partially available.

The main contributions of this work are as follows: First, we propose the new Data Science problem of Ranking of Experts' by their unobserved decision Quality (REQ). Second, we develop a machine learning-based approach for addressing the REQ problem, and which can be applied with an arbitrary predictive induction technique that produces class probability estimates. Our approach infers PGT decisions as well as the confidence of these predictions to yield a ranking. The prediction confidence element of our approach renders it particularly robust, even when incomplete decision-related information is available, or when the decisions are characterized by low predictability, more broadly. These properties are fundamental for rendering the approach particularly useful in practice. Third, we propose a novel performance measure, which we term "Verification Rate," that assesses the economic savings of an REQ approach by the amount of human expert input needed to yield the same ranking accuracy achieved by the cost-free, REQ approach. Finally, we evaluate the efficacy of our approach empirically on a real-world

expert decision dataset and demonstrate that it either yields significantly better or otherwise comparable ranking accuracy to existing alternatives. As such, our approach constitutes a benchmark for the REQ problem. We also discuss insights on the conditions under which our approach is particularly advantageous in practice, and when it is expected to yield the most accurate results.

## Related Literature

### *Worker Evaluation and Ranking*

In modern economies, a significant portion of workers routinely process information to make decisions. Brynjolfsson and McAfee (2011) report that 60% of U.S. workers perform information processing tasks. As the portion of workers routinely making decisions rises, the need to evaluate the decisions of such workers increases.

In effect, performance evaluation is one of the most central human resources practices (Ferris et al. 2008), with significant impact on organizational performance (Huselid 1995). Such evaluations are also important because they inform a host of human resource management activities, including decisions on incentives and compensation, identifying workers that require training, personnel development, and decisions regarding staffing, promotions, and lay-offs (Armstrong and Baron 2000; Milkovich et al. 2011).

Employee ranking is one of the most common measures used in performance evaluations (Milkovich et al. 2011). Employee ranking involves "sorting" employees according to some measures of their performance. Such ranking allows organizations to distinguish between workers, and also to identify the highest and lowest performing workers (Grote 2005). Diverse manual peer- or supervisor-based evaluation techniques have been developed for this purpose, such as rating relative performance and pairwise ranking (Milkovich et al. 2011). These evaluations rely on human evaluators to assess the work of peers or subordinates, thereby incurring substantial costs. As mentioned above, in many work environments such evaluations processes may also be economically or organizationally prohibitive.

### *Automated Methods for Evaluating Labeling Correctness*

In recent years a growing stream of studies in machine learning has considered evaluating the work quality of crowd-based labelers (or annotators). Specifically, these studies focused on evaluating the quality of labels generated by crowd-sourced workers. These labels are then commonly used for training supervised learning algorithms. Crowdsourcing marketplaces are, however, known to suffer from low quality work and "spammers" (Kittur et al. 2008; Kittur et al. 2013), which results in noisy labels. Therefore assessment of the labels' quality is commonly required prior to training supervised learning algorithms. Various studies have suggested methods for "correct" label inference for data instances that undergo repeated labeling[1] (Dalvi et al. 2013; Kumar and Lease 2011; Ipeirotis et al. 2010; Rodrigues et al. 2013; Zhou et al. 2012) together with methods for selecting data instances for labeling (Ipeirotis et al. 2014, Sheng et al. 2008, Wauthier and Jordan 2011), or separate methods for inferring the correct label and labeler quality while inducing the classifier (Raykar et al. 2010). Vuurens et al. (2011) evaluated the use such methods in the presence of a substantial amount of low quality labels. Other studies (e.g., Karger et al. 2011, 2014) suggest methods to minimize the number of labeled instances while assessing workers' reliability.

These studies, however, consider settings that differ from ours in important ways. In particular, their stream of work considers evaluating label correctness and labeler quality using either repeated labeling techniques (in which multiple annotators provide labels for the same instances) or comparing labels to "gold standard" (correct) labels. Such solutions are suitable for inexpensive, crowd-based tasks. In contrast, we consider settings in which acquiring experts to assess the decisions of peer experts is economically or organizationally infeasible.

The most closely related works are by Brodley and Friedl (1999) and Dekel and Shamir (2009). Similar to the works discussed above, these studies also aim to remove labeled instances to benefit induction, but similar to our setting, they do not involve acquisition of additional labeled instances. Brodley and Friedl

[1] Repeated labeling involves multiple labelers providing the label (or dependent variable value) for each data instance

(1999) consider the problem of identifying and removing individual instances with noisy (incorrect) labels, but do not consider or suggest a method for calculating the quality of the workers who produced them. Dekel and Shamir (2009), by contrast, consider the problem of removing all instances labeled by poor quality annotators and thus assess the quality of individual labelers. Yet, Dekel and Shamir (2009) only goal was to remove poor quality workers, and their study does not aim to produce highly accurate rankings of workers. In the empirical evaluations that follow, we provide a comparison to Dekel and Shamir's method and find that our approach is indeed superior for ranking workers, particularly when prediction of the correct decisions/labels is difficult. The latter condition is fundamental to our problem because it arises when decisions are non-trivial, as is often the case when experts are called to make them, or when only a partial subset of the information used by decision makers is available for future decision assessments.

# Method

We consider multiple (K) decision-making workers $W = [W_1, \ldots, W_K]$, each of whom makes multiple routine classification decisions. Workers may be auditors who decide whether a given tax return claim is fraudulent, radiologists who decide whether an image of a patient indicates a certain malady, or physicians who decide whether to pursue a certain course of treatment for a patient at a given time. For a given worker, information about $n$ past decisions is available, such that for each decision $i \in 1 \ldots n$ the worker's decision $Y_i$ is available (e.g., the medical doctor's decision whether a patient has a tumor), along with a feature vector $X_i$ reflecting (possibly partial) information relevant to the decision (e.g., features representing the relevant X-ray image of the patient). The set of decision data available to a given worker $W_j$ is denoted $S_j = \{X_i, Y_i\}_{i=1}^{n}$. The set of decision data for all $W$ workers is denoted $S$.

Our proposed method builds on a predictive modeling framework in which workers' decisions can be viewed as labels, i.e., dependent variable values, which are potentially *noisy*. Importantly, the level of noise in a given worker's decisions is inversely proportional to the *quality* of worker's decisions. If so, our challenge is to estimate the level of noise, or to otherwise rank workers by the relative quality of their decisions. As in many real-world settings in which the correct decisions are unknown at the evaluation time, and where acquiring other experts' decisions on the same instances is not possible or otherwise undesirable -- our goal is to yield a ranking without resorting to the acquisition of true decision/labels. We thus propose a data-driven approach to accurately rank the level of noise produced by each decision-maker that is applicable to a broad range of practical settings.

Specifically, given a set of decision-making expert workers W, a model $M_j$ is induced to map a set of the decision-related features X onto workers' decisions Y, and is subsequently applied to predict the most likely correct decisions for worker $W_j$, whose decision noise we aim to assess. Henceforth we refer to these predicted decisions as "Pseudo Ground Truth" Decisions (PGT Decisions). We assess the overall quality of worker $W_j's$ decisions by computing a Decision Quality (DQ) score, derived by comparing each of $W_j's$ decisions to the predicted PGT decisions, while accounting for the confidence in each of the $M_j$'s estimated PGT decisions. Workers are subsequently ranked by their relative DQ scores.

Our approach includes three complementary components on which we elaborate below: *Training and Model Induction, Calculating the Decision Scores,* and *Rankings*.

### Training Data and Model Induction

A key challenge in ensuring that this method produces an accurate estimation of expert decision quality in practice, is the choice of training data used, and the mapping induced from the data. To predict the PGT decisions, our approach employs a mapping from X onto Y induced from the set $(S - S_j)$ — i.e., all $(x_i, y_i)$ pairs reflecting decisions made by all experts, excluding expert $Wj$ whose decision quality we aim to estimate.

To produce an accurate mapping, we draw on the benefits of ensemble models and use a majority vote-based ensemble to construct $M_j$. In particular, aggregation of predictions by multiple independent models induced from different subset of instances often yields better performance than each of the base models alone (Hastie et al., 2009). However, particularly relevant for our context is that the benefits from the ensemble are more significant if there is considerable variance across the predictions of the individual base models. To generate such variance within the ensemble forming $M_j$, our approach uses K-1 base models. Each base model $B_l$ $(l \in 1 \ldots K, l \neq j)$ is induced from training pairs $(x_i, y_i) \in S_l$ that reflect the decisions made by worker $W_l$ alone. This is different from the standard bagging approach, where each base model is induced from a bootstrapped sample drawn from all the data. Such bootstrap samples are likely to yield relatively similar models. Our approach aims to benefit from the fact that each base model's training data were been generated by a different expert, and the potential resulting variance across models reflects the differences across experts.

**Calculating the Decision Scores**

Comparing the Pseudo Ground Truth (PGT) decisions to those made by each worker can offer a mean to assess workers' decision quality. Brodley and Friedl (1999) and Dekel and Shamir (2009) successfully used comparison to pseudo ground truth to evaluate the accuracy of training data labels and eliminate low quality labels for the purpose of improving model induction. However, as we discuss and demonstrate empirically below, such inferred "ground truth" is not sufficiently refined for ranking purposes, especially when the evaluation lacks access to important decision-relevant features. Specifically, these studies consider the problem of identifying and removing instances that undermine supervised learning: Brodley and Friedl (1999) consider removing individual instances with noisy labels, whereas Dekel and Shamir (2009) remove all instances of a given "poor" labeler if they are likely to undermine induction. For the setting we consider here, however, it is necessary to infer a complete ranking of all workers, not merely identify particularly poor ones. This objective entails a more refined estimation of each worker's performance and may be considered more challenging than the settings in which PGT data have been previously used.

Importantly, the mapping $M_j$ used to infer the PGT decisions may yield many incorrect predictions for two key reasons: first, $M_j$ is induced from noisy independent variable values; second in many practical settings, $M_j$ will be induced from only a subset of the information that was used by the decision makers when they made their decisions. Thus, the features from which the $M_j$ constructs the prediction model may not include all the information needed to determine the correct choice for some decisions. Such settings arise when the information used by the worker to make a decision was not recorded in entirety, either because this is not a required procedure or because it is practically infeasible. Having potentially informative features for decision making missing in our setting inevitably undermines both the induction of $M_j$, as well as its inference (Saar-Tsechansky and Provost, 2007).

To address this challenge, we introduce the use of the confidence of the prediction and weigh the PGT decisions by the relative confidence of the corresponding predictions generated by model $M_j$. Thus, rather than account for an expert's decision $i$ as incorrect if it merely differs from the decision predicted by model $M_j$, we scale this quality judgment by the *confidence* of model $Mj$'s prediction for decision $i$. Thus, the estimation of an expert's decision accuracy is informed more heavily by model $M_j$'s most confident predictions of the correct decision, and less so by $M_j$'s least confident predictions. Formally, $S_j$ denotes the set of all decisions made by worker $j$. Let $s_j^+ \subset S_j$ denote the set of all decisions made by worker $j$ for which model $M_j$ predicts the same decision, and $s_j^- \subset S_j$ denote the set of all decisions by worker $j$ with which model $M_j$ disagrees (predicts a different decision than worker $W_j$). The decision quality score for worker $j$ is then given by:

$$DQ(W_j) = \left( \frac{\sum_{i \in s_j^+} conf_i}{\sum_{i \in S_j} conf_i} \right)$$

where $Conf_i$ refers to the confidence of model $M_j$ in its prediction for decision $i$. We capture model $M_j's$ confidence for decision $i$ by the proportion of base models ($B_\iota s$ – individual models within the ensemble) that constitute the majority vote in favor of ensemble model $M_j$'s prediction for decision $i$. (For example, if only 55% of the ensemble base models agree on the choice for predicting decision $i$ then $Conf_i$ = 0.55).

Accounting for model $M_j's$ prediction confidence becomes increasingly important the less accurate model $M_j's$ predictions are. As mention above, poor predictions may arise when the feature set $X$ describing the decision does not capture sufficient information to induce an accurate model. This may be indeed the case in many practical settings in which workers had access to a much richer information set when making the decision than the information available at the time one wishes to retrospectively evaluate that decision. In the empirical evaluations that follow, we find empirically the significant effect of the confidence scaling when important information for the decisions is absent from the feature set $X$ and when overall predictability is low.

### Ranking

Finally, once the quality $DQ(W_j)$ of each individual worker $W_j$ is determined, workers are straightforwardly ranked by their respective $DQ(W_j)$ scores.

## Experimental Setup

Following common practice in the large stream of literature dealing with labeling data quality (For example, Ipeirotis et al., 2014; Raykar et al. 2010; Sheng et al. 2008), we use simulation to evaluate performance. Simulation allows us to maintain consistent conditions, while conducting multiple replications of the same experiment. Furthermore, simulation allows setting up challenging conditions in which the workers decision quality only slightly varies across different workers.

Figure 1 provides an overview of the simulation procedure. The simulation process begins by taking in a dataset that includes a large number of instances. Each instance has a set of features (predictors) and the actual decision made (the binary class label). The simulation procedure then randomly splits the dataset and assigns different data instances to 20 simulated workers.[2] Subsequently, the simulation procedure assigns one of several predetermined levels of quality to each worker. Quality levels are reflected by a worker's probability to make a correct decision. Based on this probability, the simulation procedure "injects" decision errors into the worker's decisions. For example, if a worker is assigned a quality level of 95% (probability to make a correct decision), for each instance assigned to this worker, the simulation draws a random number and if it exceeds 0.95 the simulation procedure changes the original label into the opposite label. This process is repeated for all workers according to their assigned probabilities.

Implementing our simulation, we intentionally use a challenging setting in which decision quality only slightly varies across workers – making it difficult for automated methods to detect differences across workers. Specifically, we introduce differences in decision accuracy of only 1% between the workers. The top-performing worker is assigned a decision accuracy of 100%, the second best worker is assigned a decision accuracy of 99% and so on, with the least accurate of the 20 workers assigned a decision accuracy of 81%. Our method is applied only after the decision errors are injected into the labels, with the goal of discovering the original rankings of the workers. In this simulation, we implement our method using a standard Random Forest algorithm (R package RandomForest with default settings and 1000 sub trees). This simulation procedure is repeated 50 times.

Finally, it is important to note that the original ranking, and the specific instances to which the simulation injected decision inaccuracies, remain strictly hidden from our method and the only way our method can uncover this information is directly from the data itself.

[2] Each worker is assigned 1/20 of the instances in the dataset.

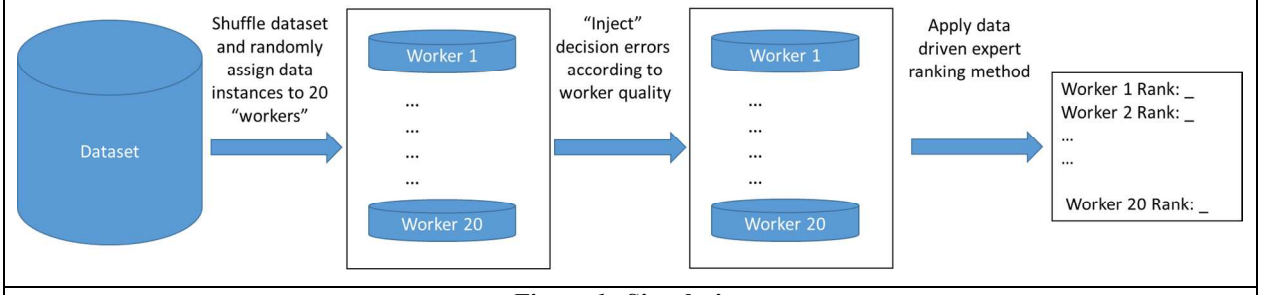


***Figure 1.*** **Simulation**

Figure 1 outlines a single repetition of the simulation procedure. The results that we report are based on repeating this procedure 50 times.

### *Performance Measures*

After running each simulation replication we measure how well our method re-constructs the original, simulated, worker ranking. This is achieved by comparing the ranking of our method to the simulated ranking (based on which we injected different decision quality levels). Specifically, we use two well-known performance measures that capture the correlation between two sets of ranking: Spearman's Rho (also known as Spearman's rank correlation coefficient) and Kendall's tau (Kendall rank correlation coefficient). Both performance measures have the range of +1 to -1 (With +1 value indicating that the original set of ranking and our set of ranking are most positively correlated).

In addition to these well-known ranking measures, we also suggest a novel performance measure that is designed to assess the economic value, or the amount of work that our method can save organizations.[3] We call this measure *Verification Rate*.

This measure emulates an alternative decision evaluation process by an oracle (i.e., an expert or a committee of experts with 100% accuracy in their judgments), which reviews a sample of the decisions made by each worker, manually assesses their correctness or incorrectness, and then ranks the workers according to the proportion of the correct decisions they made, based on the limited sample reviewed by the oracle. Naturally, the higher the sampling rate used by the oracle, the more accurate the oracle's ranking (although more work would also be required).

Therefore, for any ranking accuracy obtained by our method, the *Verification Rate* captures the proportion of each worker's decisions that an oracle must manually assess to achieve the same level of accuracy. Specifically, for each experiment we first calculate the Spearman Rho measure obtained by our method. We then report the sampling ratio (percentage of instances out of each worker's data instances) required to produce the same Spearman Rho as the Verification Rate.

This sampling ratio is calculated empirically in the following manner: We first run our method and record *SR* - the Spearman Rho our method obtains. We then, independent of our method's ranking, begin sampling a small ratio (0.1%) of each worker's decisions. After obtaining this small sample we consider only the decisions included in this sample, verify their correctness, and measure the percent of correct decisions for each worker. We then rank the workers according to the percentage of their correct decisions in this sample and calculate the Spearman Rho. (a very low sampling ratio typically results in a low Spearman Rho value since most of the decisions are not evaluated). We then repeat this process while continuously increasing the sampling ratio (by 0.1% increments) until the sampling ratio is sufficient to produce a Spearman Rho that is equivalent to the *SR* value our method obtained. We report this final sampling ratio as the *Verification Rate*. The higher the Verification Rate, the greater the amount of manual work and costs that our method could save organizations.

[3] While we note that a major motivation for this research is the fact that in various organizational settings peer/supervisor based ranking is infeasible – it is still interesting to quantify the amount of work our method could save. Similar to the Spearman Rho and Kandell's tau measures, Verification Rate is measured only *after* the ranking is performed, and is not used during the ranking process.

### *Datasets*

Our primary data set for evaluating our methodology corresponds to about 30,000 decisions made by sales tax auditors to determine whether or not companies in the State of Texas have complied with the state tax law. Note that the information available to the human auditors included all the business records available in each firm, while the predictors provided in this dataset includes merely summaries of the firm's information, such as its age, stated sales relative to its comparable competitors and wages, as well as prior audit history, such as prior positive audits (which produced revenues to the state), etc.

The Tax Audits dataset provides a challenging, real-world setting for evaluating our method. The dataset includes real human decisions made by expert tax auditors. Furthermore, the predictors (independent variables) differ from the potentially rich information set available to the tax auditors during an audit. For example, the tax auditors can inspect complete financial records, study individual transactions, ask firm officials for clarifications, etc. Such differences in the information set available to decision makers but not for the predictive models used to predict PGT decisions is a likely characteristic of many practical settings. As discussed, this lack of information is likely to result in weak predictive models generating the PGT decisions. It is therefore interesting to explore whether our approach method can yield effective ranking in such a setting.

Finally, it is important to note that for this dataset we were unable to obtain access to the details (such as ID) of the original tax auditors. Therefore, our simulation procedure treats the dataset (as described in the simulation section above) by randomly allocating audit decisions to different workers whose simulated level of quality is represented by the injected decisions errors.

We note that while we treat the original decisions in this dataset as (absolute) ground truth for the purpose of the simulation, the original decisions may also contain some errors. Yet, since in our data the instances are randomly split across the different simulated workers, we do not expect such errors to consistently bias our results in favor of a certain simulated scenario or worker. In the worst case, any inaccuracies in the original dataset can be regarded as additional noise in the data that is likely to confound our method. Thus, the results we report here for our method could be regarded as conservative estimates.

For robustness, we also simulate and test our method on two well-known publicly available datasets: The Mushroom and SPAM datasets (Lichman 2013). The Mushroom dataset includes expert decisions on whether mushrooms are edible or not. The SPAM dataset includes binary decisions whether emails are SPAM. Besides providing additional robustness, both Mushroom and SPAM datasets are, unlike the Tax Audits data, characterized by a high level of predictability and relatively low noise. Therefore, using our three datasets, we can explore the efficacy of our method under varying levels of predictability (As a measure to the level of inherent predictability in each dataset, we measured AUC when running a 10-fold cross validation procedure. The AUC for the Tax Audit dataset was a relatively low 0.67, AUC for the SPAM dataset was high - 0.987; and the Mushroom dataset had a perfect AUC of 1.0).

## Results

### *Tax Audits Dataset*

Our primary set of results relates to the Tax Audit task. As mentioned above, the Tax Auditors ranking task is based on real-life data and has several challenging characteristics, including the fact that the explanatory variables available for our method do not include all the information that was available to the auditors, and the fact that the dataset itself displays low predictability. Consequently, we could expect a significant amount of noise when generating PGT decisions by our method. Given these challenging characteristics, it is interesting to evaluate the performance of our method.

Table 1 presents average ranking statistics and average verification rates for 50 experiments using the Tax Audit data. It is important to note that in each of the 50 experiments both Spearman's Rho and Kendall's tau statistics were found to be statistically significant with a P-Value lower than 0.01 (detailed, per experiment, results are not presented due to space constraints, but are available upon request from the authors). Furthermore, we observe that the average Verification Rate was 2% (~600 instances out of ~30,000 instances), which implies that a manual review of 2% data by a highly trained expert or committee of experts (who are able to produce 100% correct judgments) would be necessary to obtain a similar level of ranking accuracy. Such a manual evaluation task would impose significant labor costs. This result especially stands out as our ranking method would generate a similar level of ranking accuracy for free.

| Average Spearman's Rho | Average Kendall's tau | Average Verification Rate |
|---|---|---|
| 82.9% | 65.4% | 2% (~600 decisions) |

**Table 1 – Performance Measure for Ranking Tax Audits Decision Makers**

To provide additional insights about the capabilities of our method, Table 2 presents a confusion matrix for the ranking of the tax audit experts. Based on 50 experiments, this table shows the percentage of experiments in which the $i_{th}$ worker is ranked in terms of decision accuracy (vertical axis) as the $j_{th}$ worker by our model (horizontal axis). For example, in the upper left-hand corner of Table 2, we observe that the 1st worker (top performing worker) was ranked by our model as the 1st worker in 40% of the experiments, and was ranked as the 2nd best worker in 12% of the experiments. Overall results show that our method is especially useful for identifying the least accurate decision makers and the most accurate decision makers. For example, there is an 80% chance that the least accurate worker (worker ranked number 20) will be included in the list of the 5 lowest ranked workers by our model. As identifying low-performing workers is a key task in many settings, such as medical or security applications, our method can be used to suggest a small set of candidates whose work warrants careful monitoring.

In summary, considering the challenging characteristics of the Tax Audit dataset, which includes missing information compared to the original decision makers' information and low predictability within the dataset, our method generates an overall accurate ranking that is statistically significant, provides substantial economic savings (as measured by the Verification Rate), and also effectively identifies low and high performing workers.

| | | Model-based Decision Ranking (1- highest, 20- lowest) | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Actual Rank (probability of a correct decision) 1- highest 20 - lowest | 1 | 40% | 12% | 8% | 12% | 4% | 8% | 8% | 6% | | 2% | | | | | | | | | | |
| | 2 | 22% | 36% | 14% | 10% | 6% | 6% | 2% | 2% | | 2% | | | | | | | | | | |
| | 3 | 14% | 14% | 22% | 8% | 12% | 10% | 4% | 2% | 4% | 4% | 2% | 2% | 2% | | | | | | | |
| | 4 | 14% | 20% | 16% | 16% | 10% | 2% | 2% | 10% | 4% | 2% | 2% | 2% | | | | | | | | |
| | 5 | 4% | 8% | 14% | 14% | 14% | 18% | 4% | 4% | 4% | 8% | 4% | | | | 2% | 2% | | | | |
| | 6 | 4% | | 14% | 12% | 10% | 14% | 12% | 6% | 10% | 8% | 2% | 2% | 2% | 2% | 2% | | | | | |
| | 7 | | 2% | 4% | 10% | 8% | 6% | 14% | 18% | 4% | 8% | 6% | 12% | 4% | 2% | 2% | | | | | |
| | 8 | 2% | 2% | 6% | 4% | 10% | 8% | 16% | 12% | 10% | 8% | 8% | 4% | | 6% | 2% | 2% | | | | |
| | 9 | | 2% | 2% | 2% | 10% | 6% | 16% | 12% | 10% | 10% | 4% | 8% | 6% | 4% | 6% | | | 2% | | |
| | 10 | | 4% | | 6% | 6% | 6% | | 8% | 8% | 10% | 12% | 8% | 8% | 10% | 2% | 2% | 2% | 6% | 2% | |
| | 11 | | | | 2% | 4% | 4% | 8% | 10% | 12% | 8% | 8% | 10% | 10% | 6% | 8% | 4% | 6% | | | |
| | 12 | | | | 2% | 2% | 6% | 6% | 4% | 6% | 2% | 10% | 18% | 12% | 6% | 10% | 2% | 4% | 2% | 4% | 4% |
| | 13 | | | | | 2% | 2% | 2% | 2% | 14% | 14% | 12% | 14% | 6% | 6% | 2% | 8% | 8% | | 2% | 6% |
| | 14 | | | | | | | 6% | 4% | | 4% | 6% | 4% | 10% | 16% | 8% | 14% | 14% | 4% | 4% | 6% |
| | 15 | | | | 2% | | | | | 2% | 4% | 2% | 4% | 12% | 10% | 18% | 8% | 10% | 14% | 10% | 4% |
| | 16 | | | | | | 2% | | | 8% | 4% | 4% | 4% | 6% | 4% | 18% | 18% | 6% | 16% | 8% | 2% |
| | 17 | | | | | 2% | | | | 2% | | 8% | | 10% | 8% | 4% | 12% | 16% | 14% | 12% | 12% |
| | 18 | | | | | | 2% | | | | 2% | 6% | 6% | 2% | 10% | 10% | 10% | 12% | 16% | 20% | 4% |
| | 19 | | | | | | | | | 2% | | 4% | | 6% | 2% | | 8% | 10% | 16% | 14% | 38% |
| | 20 | | | | | | | | | | | | 2% | 4% | 8% | 6% | 10% | 12% | 10% | 24% | 24% |

Table 2 - Results are averages over 50 experiments based on random data splits. Blank cells represent zero value. Darker colors represent higher rates.

**Table 2 – Confusion Matrix and Heat Map for Ranking Tax Audits Decision Makers**

### *Mushroom Dataset*

For robustness, we evaluated the performance of our method using a second dataset, the well-known Mushroom dataset. This dataset was initially labelled by experts. Therefore, introducing noise into the expert decisions, as discussed in the "Experimental Setup" section would also simulate real-life decision errors of expert workers. Unlike the Tax Audit data, this dataset is characterized by very high predictability.

As detailed in Table 3, the high predictability of this dataset translates into improved results: Our method obtains very high Spearman's Rho and Kendall's tau values. Similar to the Tax Audit data, both Spearman's Rho and Kendall's tau statistics in all 50 experiments were found to be statistically significant, with a P-Value lower than 0.01. Moreover, due to the very accurate ranking of our method, the Verification Rate is very high (83.6%), suggesting substantial economic savings.

| Average Spearman's Rho | Average Kendall's tau | Average Verification Rate |
|---|---|---|
| 99.9% | 99.1% | 83.6% (~6,800 decisions) |

**Table 3 – Performance Measure for Ranking Mushroom Classification Decisions**

The confusion matrix presented in Table 4 furthermore demonstrates that our method was successful in generating an accurate individual-level ranking, and achieved an especially high level of accuracy when ranking the highest and lowest performing expert workers. For example, in 100% of the experiments the most accurate decision maker was ranked as such; and in 98% of the experiments, the least accurate worker was indeed ranked as the least accurate worker.

| | | Model-based Decision Ranking (1- highest, 20- lowest) | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Actual Rank (probability of a correct decision) 1- highest 20 - lowest | 1 | 100% | | | | | | | | | | | | | | | | | | | |
| | 2 | | 100% | | | | | | | | | | | | | | | | | | |
| | 3 | | | 94% | 6% | | | | | | | | | | | | | | | | |
| | 4 | | | 6% | 94% | | | | | | | | | | | | | | | | |
| | 5 | | | | | 94% | 6% | | | | | | | | | | | | | | |
| | 6 | | | | | 6% | 94% | | | | | | | | | | | | | | |
| | 7 | | | | | | | 94% | 6% | | | | | | | | | | | | |
| | 8 | | | | | | | 6% | 92% | 2% | | | | | | | | | | | |
| | 9 | | | | | | | | 2% | 90% | 8% | | | | | | | | | | |
| | 10 | | | | | | | | | 8% | 86% | 6% | | | | | | | | | |
| | 11 | | | | | | | | | | 6% | 90% | 4% | | | | | | | | |
| | 12 | | | | | | | | | | | 4% | 88% | 8% | | | | | | | |
| | 13 | | | | | | | | | | | | 8% | 90% | 2% | | | | | | |
| | 14 | | | | | | | | | | | | | | 92% | 6% | 2% | | | | |
| | 15 | | | | | | | | | | | | | 2% | 6% | 82% | 10% | | | | |
| | 16 | | | | | | | | | | | | | | | 12% | 84% | 4% | | | |
| | 17 | | | | | | | | | | | | | | | | 4% | 90% | 6% | | |
| | 18 | | | | | | | | | | | | | | | | | 6% | 86% | 8% | |
| | 19 | | | | | | | | | | | | | | | | | | 8% | 90% | 2% |
| | 20 | | | | | | | | | | | | | | | | | | | 2% | 98% |

Note. Results are averages over 50 experiments based on random data splits. Blank cells represent zero value. Darker colors represent higher rates.

**Table 4 – Confusion Matrix and Heat Map for Ranking Mushroom Classification Decisions**

### *SPAM Dataset*

For additional robustness, we evaluated our method using a third dataset — the well-known SPAM dataset. The the SPAM dataset is also characterized by very high predictability. As detailed in Table 5, this high predictability translates into improved results for our method (high Spearman's Rho and Kendall's tau values). Similar to the other datasets -- both Spearman's Rho and Kendall's tau statistics were found to be statistically significant with a P-Value lower than 0.01 in all 50 experiments. As before, due to the accurate ranking of our method – the Verification Rate is relatively high, indicating substantial savings. The confusion matrix presented in Table 6 also presents a very accurate identification of the highest- and lowest-performing expert workers, similarly to the previous tasks.

In sum, both publicly available datasets -- the Mushroom and SPAM datasets -- provide additional evidence on the efficacy of our method. Moreover, the improved performance of our method when applied on the Mushroom and Spam datasets, which have high levels of inherent predictability than the Tax Audit data offers evidence of the conditions in which our method obtains better results. Specifically, we observe that our method can produce even more accurate rankings when the predictability in the data is high.

| Average Spearman's Rho | Average Kendall's tau | Average Verification Rate |
|---|---|---|
| 94.4% | 82.5% | 37% (~1700 decisions) |

**Table 5 – Performance Measure for Ranking SPAM Decisions**

| | | Model-based Decision Ranking (1- highest, 20- lowest) | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Actual Rank (probability of a correct decision) 1- highest 20 - lowest | 1 | 50% | 30% | 10% | 6% | 2% | | 2% | | | | | | | | | | | | | |
| | 2 | 28% | 34% | 14% | 16% | 2% | 6% | | | | | | | | | | | | | | |
| | 3 | 12% | 10% | 36% | 16% | 8% | 10% | 4% | 2% | 2% | | | | | | | | | | | |
| | 4 | 4% | 20% | 14% | 18% | 20% | 8% | 8% | 6% | 2% | | | | | | | | | | | |
| | 5 | 4% | 4% | 12% | 22% | 24% | 12% | 6% | 4% | 8% | 4% | | | | | | | | | | |
| | 6 | 2% | 2% | 2% | 6% | 22% | 18% | 18% | 16% | 6% | 4% | 2% | 2% | | | | | | | | |
| | 7 | | | 10% | 8% | 12% | 22% | 16% | 14% | 10% | 4% | 4% | | | | | | | | | |
| | 8 | | | 2% | 4% | 4% | 14% | 20% | 20% | 14% | 14% | 2% | 4% | | 2% | | | | | | |
| | 9 | | | | 4% | 2% | 6% | 10% | 24% | 14% | 20% | 12% | 4% | 4% | | | | | | | |
| | 10 | | | | | 2% | 2% | 12% | 2% | 22% | 22% | 20% | 12% | 4% | 2% | | | | | | |
| | 11 | | | | | 2% | | 4% | 12% | 12% | 16% | 22% | 18% | 8% | | 6% | | | | | |
| | 12 | | | | | | 2% | | | 6% | 4% | 14% | 24% | 34% | 4% | 6% | 4% | | 2% | | |
| | 13 | | | | | | | | | 2% | 8% | 8% | 22% | 8% | 30% | 14% | 6% | 2% | | | |
| | 14 | | | | | | | | | 2% | 2% | 8% | 2% | 16% | 28% | 20% | 16% | 4% | | | 2% |
| | 15 | | | | | | | | | | 2% | 2% | 8% | 16% | 14% | 20% | 14% | 8% | 8% | 6% | 2% |
| | 16 | | | | | | | | | | | 4% | | 6% | 10% | 14% | 30% | 20% | 6% | 8% | 2% |
| | 17 | | | | | | | | | | | | 4% | 2% | 6% | 12% | 10% | 30% | 24% | 10% | 2% |
| | 18 | | | | | | | | | | | | | 2% | 2% | 4% | 12% | 24% | 28% | 16% | 12% |
| | 19 | | | | | | | | | | | | | | 2% | 4% | 4% | 8% | 20% | 40% | 22% |
| | 20 | | | | | | | | | | | 2% | | | | | 4% | 4% | 12% | 20% | 58% |

Note: Results are averages over 50 experiments based on random data splits. Blank cells represent zero value. Darker colors represent higher rates.

**Table 6 – Confusion Matrix and Heat Map for Ranking SPAM Classification Decisions**

## *Comparison to an Alternative Approach*

As discussed in the literature review, Dekel and Shamir (2009) proposed an approach that was designed to address a different problem. Nevertheless, since Dekel and Shamir's method also generates worker PGT scores, it is interesting to evaluate how our method ranking compares to a ranking generated when the decisions are scored using Dekel and Shamir's method. We compare the performance of the two methods in Table 7.

**Table 7A**
**Average Spearman's Rho (50 repetitions)**

| Dataset | REQ | D & S | Improvement using REQ |
|---|---|---|---|
| Tax Audit | 82.9% | 78.8% | 4.1%*** |
| Mushroom | 99.9% | 99.9% | 0.0% |
| SPAM | 94.4% | 93.5% | 0.9%** |

**Table 7B**
**Average Kendall's tau (50 repetitions)**

| Dataset | REQ | D & S | Improvement using REQ |
|---|---|---|---|
| Tax Audit | 65.4% | 60.4% | 5.0%*** |
| Mushroom | 99.1% | 99.5% | -0.4% |
| SPAM | 82.5% | 81.1% | 1.4%** |

Table 7A (left) reports average spearman Rho and Table 7B (right) report average Kendall's tau. REQ column presents the results for our method whereas D & S column presents the results obtained by Dekel and Shamir's (2009) baseline method. We report the significance of the improvement in performance using bootstrap P-Value: * denotes P-values lower than 0.1; **denotes P-values lower than 0.05; *** denotes P-values lower than 0.01.

**Table 7A &B – Comparison between our Method (REQ) and Baseline Method**

These results show that our method obtained equivalent or better ranking compared to those obtained using Dekel and Shamir's baseline method, for all the datasets. Moreover, our method performs significantly better in the more challenging case of the Tax Audit dataset, where we expect the PGT decisions to be of lower quality. This finding is in line with the expectations from our method that is designed to handle cases of lower PGT decision quality, by taking into account the method's confidence in each PGT decision. In the case of a less challenging dataset, such as SPAM, we see that the advantage of using our method is more limited. Finally, in the case in which the dataset displays perfect predictability,

such as in case of the Mushroom dataset, both our method and the baseline method, generate near perfect results.

### *Simulating Missing Information*

As mentioned, one of the major challenges of REQ problem is the fact that in real-life settings, a portion of the information that was available to the decision maker may not be available when the decision is being evaluated. For example, in some cases tacit information that was available to the decision maker is not available at the time of ranking, or not all of the original information was recorded. As a robustness check of whether our method design can produce an accurate ranking under missing information, we conducted the following experiment with the Mushroom and SPAM datasets, which both had very high initial rates of predictability.[4] Specifically, to simulate missing information, we remove the first 90% of the features in each of the datasets.

Table 8 summarizes our method's performance and includes a comparison to Dekel and Shamir's baseline method. We note that although the performance of our method unsurprisingly was lower compared to when the full set of information was available, our method nonetheless obtains very good ranking in terms of Spearman's Rho and Kendall's tau statistics, which were again statistically significant in all 50 experiments.

Furthermore, comparing Tables 7 and 8, we observe that in case of missing information (Table 8), the improvement obtained when using our method in comparison to Dekel and Shamir's baseline method is now much greater compared to when the evaluation was conducted on the original datasets with no missing features (Table 7). For example, after removing 90% of the features, the improvement in Pearson Rho that our method obtain compared to the baseline Dekel and Shamir method was 2.5% for the Mushroom dataset (table 8) compared to 0% improvement (Table 7) when the Mushroom dataset included all features. We observe a similar increase in improvement compared to the baseline method for the SPAM dataset as well and when considering Kendall's tau statics in both datasets. These results offer additional evidence about the usefulness of our method in practical cases with no access to the full set of information that was available to the original decision maker, and in cases of low predictability in the data. In such conditions, which lead to lower quality PGT decisions, our confidence weighting mechanism provides an advantage

**Table 8A**

**Average Spearman's Rho (50 repetitions)**

| Dataset | REQ | D & S | Improvement using REQ |
|---|---|---|---|
| Mushroom | 88.0% | 85.5% | 2.5%*** |
| SPAM | 90.0% | 88.1% | 2.0%*** |

**Table 8B**

**Average Kendall's tau (50 repetitions)**

| Dataset | REQ | D & S | Improvement using REQ |
|---|---|---|---|
| Mushroom | 71.8% | 68.8% | 3.0%*** |
| SPAM | 75.4% | 72.4% | 2.9%*** |

Table 8A (left) reports average spearman Rho and Table 8B (right) reports average Kendall's tau. REQ column details the results for our method whereas D & S column details the results obtained by Dekel and Shamir's (2009) baseline method. We report the significance of the improvement in performance using bootstrap P-Value: * denotes P-values lower than 0.1; **denotes P-values lower than 0.05; *** denotes P-values lower than 0.01.

**Table 8 A&B – Comparison between our Method (REQ) and Baseline Method (Data is Missing 90% of the Predictors)**

### *Effect of the Confidence Mechanism*

As mentioned above, a key feature in our methodology is the confidence mechanism designed to give less weight to measured decision errors in cases where we have lower confidence in the accuracy of the PGT decisions. So far, we provided indirect evidence to the efficacy of this mechanism in comparison with a baseline ranking method. Nevertheless, in this sub-section we provide a more direct evaluation of this feature. Specifically Table 9 shows the effect of "turning-off" this feature by comparing the performance of our method with a second, similar, method. The only difference between the two methods is that in the

[4] We do not repeat this analysis using the Tax Audit data since we know it already does not include important information that was available to the auditors.

second method – all measured errors receive an equal weight. As evident from this table there is a substantial difference in performance when activating or deactivating the confidence weighting mechanism thus providing direct support to the usefulness of this feature. The only exception is in case of the mushroom dataset which is characterized by unusually high predictability. This result is expected as the PGT decisions weighting scheme was introduced to handle cases in which the confidence in the predicted PGT decisions is not very high. Yet, when the data presents very high predictability the confidence weighting scheme is not needed for improving the results.

**Table 9A**
**Average Spearman's Rho (50 repetitions)**

| Dataset | With Confidence Weighting | Without Confidence Weighting | Improvement by Activating Weighting Feature |
|---|---|---|---|
| Tax Audit | 82.9% | 79.0% | 3.9%*** |
| Mushroom | 99.9% | 99.8% | 0.1% |
| SPAM | 94.4% | 93.0% | 1.4%*** |

**Table 9B**
**Average Kendall's tau (50 repetitions)**

| Dataset | With Confidence Weighting | Without Confidence Weighting | Improvement by Activating Weighting Feature |
|---|---|---|---|
| Tax Audit | 65.4% | 61.4% | 4.1%*** |
| Mushroom | 99.1% | 98.9% | 0.2%* |
| SPAM | 82.5% | 80.2% | 2.3%*** |

Table 9A (left) reports average spearman Rho and Table 9B (right) report average Kendall's tau. "With Confidence Weighting" column details the results for our standard method whereas "Without Confidence Weighting" column details the results when our method confidence weighting feature is deactivated. We report the significance of the improvement in performance using bootstrap P-Value: * denotes P-values lower than 0.1; **denotes P-values lower than 0.05; *** denotes P-values lower than 0.01.

**Table 9A&B – Comparison of our Method Performance with and without the Confidence Weighting Feature**

## Summary and Conclusions

The widespread use of Data Science and Business Analytics methods to improve business performance has increased in a host of domains in recent years. In this paper, we suggest a solution to a novel Data Science problem – the REQ problem: Ranking Expert decision-makers' unobserved Quality. To overcome the challenge of ranking when true quality is unobserved we propose a new method that both generates PGT decisions and assesses their reliability. Testing on three datasets, we find that the method is successful in generating accurate and statistically significant expert worker rankings and that it obtains superior or comparable results compared to a baseline method. Moreover, using a novel performance measure, Verification Rate, we show that the economic value, represented by the amount of manual work that our method could save, is substantial.

As demonstrated in this paper, the unique feature of our method -- assessments of the confidence in PGT decisions – underlies our method's improved performance in realistic and challenging conditions in which the recorded information (or predictors) available for modeling may be merely a small subset of the information that was originally available to the decision makers, or in cases that overall predictability is low.

Finally, in this paper we also discuss the conditions in which our method obtains best performance. In general, our method achieves higher performance levels when predictability in the data is high, and when no predictors are omitted. Nevertheless, in the challenging cases of low predictability in the data, our method's weighting feature gives our method additional robustness, and in these challenging conditions, our method shows an even greater improvement compared to the baseline method.

### *Business and Managerial Implications*

Assessment of workers' performance is a key managerial function. It is used for a host of human resource management activities such as identifying workers who need training, designing worker incentives and compensation, as well as staffing decisions. As worker decision quality is unobservable in many settings, our method has substantial business implications, as it allows ranking of workers in cases where such ranking is organizationally or economically prohibitive. Moreover, using the Verification Rate measure,

we show that our method can be used to create substantial savings, even in settings in which manual evaluation of past decisions is possible.

***Limitations and Future Work***

In this work we propose a new method to address the REQ problem. Although this method produces good results additional research can further improve the generation of PGT decisions. One possible research direction would be to give different weights, in an iterative manner, to each worker-based classifier within the ensemble of classifiers according to the individual worker's accuracy: after worker's quality is estimated, the classifier trained on each worker's data would then be weighted according to the worker's reliability. Then, worker quality would be assessed again, and so on. Another approach might involve a series of distinct inducers that are trained on the entire data rather than on individual workers. Although in this study we have so far considered only classification decisions, it is certainly possible to develop an extension to our method to address other types of expert decisions. Finally, in this paper we considered settings where workers encounter similar distribution of decisions (e.g., similar distribution of decision difficulty, etc.). However, in some specialty areas experts may draw tasks of different nature and require methods that would bring to bear the difficulty of the tasks in order to assess experts' relative decision-quality.

## References


Armstrong, M., and, Baron, A. 2000. "Performance Management," in *Human Resource Management*, R. Dransfield (ed.), Oxford, UK: Heinemann, pp. 69-84.

Brodley, C. E., and Friedl, M. A. 1999. "Identifying Mislabeled Training Data," *Journal of Artificial Intelligence Research* (11), pp. 131-167.

Brynjolfsson, E., and McAfee, A. 2011. *Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy*. Lexington, MA: Digital Frontier.

Dalvi, N., Dasgupta, A., Kumar, R., and Rastogi, V. 2013. "Aggregating Crowdsourced Binary Ratings," in *Proceedings of the 22nd International Conference on World Wide Web*, (pp. 285-294).

Dekel, O., and Shamir, O. 2009. "Vox Populi: Collecting High-Quality Labels from a Crowd," in *Proceedings of the 22nd Annual Conference on Learning Theory*.

Ferris, G. R., Munyon, T. P., Basik, K., and Buckley, M. R. 2008. "The Performance Evaluation Context: Social, Emotional, Cognitive, Political, And Relationship Components," Human Resource Management Review (18:3), pp. 146-163.

Grote, R. C. 2005. *Forced Ranking: Making Performance Management Work*. Boston, MA: Harvard Business Press.

Hastie, T., R. Tibshirani, and J. Friedman. "The elements of statistical learning 2nd edition.". New York, Springer (2009).

Huselid, M. A. 1995. "The Impact of Human Resource Management Practices on Turnover, Productivity, and Corporate Financial Performance.," *Academy of Management Journal* (38:3), pp. 635-672.

Ipeirotis, P. G., Provost, F., and Wang, J. 2010. "Quality Management on Amazon Mechanical Turk," in *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pp. 64-67.

Ipeirotis, P. G., Provost, F., Sheng, V. S., and Wang, J., 2014. "Repeated Labeling Using Multiple Noisy Labelers,". *Data Mining and Knowledge Discovery*, (28:2), pp. 402-441.

Karger, D. R., Oh, S., and Shah, D. 2011. "Iterative Learning For Reliable Crowdsourcing Systems," in *Proceedings of Advances in Neural Information Processing Systems: 25th Annual Conference on Neural Information Processing*.

Karger, D. R., Oh, S., and Shah, D. 2014. "Budget-Optimal Task Allocation for Reliable Crowdsourcing Systems," . *Operations Research* (62:1), February, pp. 1-24.

Kittur, A., Chi, E. H., and Suh, B. 2008. "Crowdsourcing User Studies with Mechanical Turk," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., and Horton, J. 2013. , "The Future of Crowd Work," in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work,* pp. 1301-1318.

Kumar, A., and Lease, M. 2011. "Modeling Annotator Accuracies for Supervised Learning," in *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM)*, at the Fourth ACM International Conference on Web Search and Data Mining (WSDM), pp. 19-22.

Lichman, M. 2013. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Milkovich, G. T., and Newman, J. M. 2011. "*Compensation*. "New York, NY: McGraw-Hill.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. 2010. "Learning From Crowds," . *The Journal of Machine Learning Research* (1:11), March, pp. 1297-1322.

Rodrigues, F., Pereira, F., and Ribeiro, B. 2013. "Learning from Multiple Annotators: Distinguishing Good from Random Labelers," *Pattern Recognition Letters* (34:12), September, pp. 1428-1436.

Saar-Tsechansky, M., and Provost, F. 2007. "Handling Missing Values When Applying Classification Models," *Journal of Machine Learning Research* (8), July, pp. 1623-1657.

Sheng, V. S., Provost, F., and Ipeirotis, P. G., 2008. "Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 614-622.

Vuurens, J., de Vries, A. P., and Eickhoff, C. 2011. "How *Much Spam Can You Take? An Analysis of Crowdsourcing Results to Increase Accuracy," in Proceedings of ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)*, pp. 21-26.

Wauthier, F. L., and Jordan, M. I. 2011. "Bayesian Bias Mitigation for Crowdsourcing," in *Advances in Neural Information Processing Systems (NIPS)*, P. Bartlett, F. Pereira, J. Shawe-Taylor and R. Zemel (eds.), pp. 1800-1808.

Zhou, D., Basu, S., Mao, Y., and Platt, J. C. 2012. "Learning from the Wisdom of Crowds by Minimax Entropy," In *Advances in Neural Information Processing Systems*, pp. 2204-2212.